scientific reports

Check for updates

OPEN

Information extraction from historical well records using a large language model

Zhiwei Ma[1]✉, Javier E. Santos[1], Greg Lackey[2], Hari Viswanathan[1] & Daniel O'Malley[1]

To reduce environmental risks and impacts from orphaned wells (abandoned oil and gas wells), it is essential to first locate and then plug these wells. Manual reading and digitizing of information from historical documents is not feasible, given the large number of wells. Here, we propose a new computational approach for rapidly and cost-effectively characterizing these wells. Specifically, we leverage the advanced capabilities of large language models (LLMs) to extract vital information including well location and depth from historical records of orphaned wells. In this paper, we present an information extraction workflow based on open-source Llama 2 models and test it on a dataset of 160 well documents. The developed workflow achieves an overall accuracy of 100%, accounting for both text conversion and LLM analysis when applied to clean, PDF-based reports. However, it struggles with unstructured image-based well records, where accuracy drops to 70%. The workflow provides significant benefits over manual human digitization, because it reduces labor and increases automation. Additionally, more detailed prompting leads to improved information extraction, and LLMs with more parameters typically perform better. Given that a vast amount of geoscientific information is locked up in old documents, this work demonstrates that recent breakthroughs in LLMs allow us to access and utilize this information more effectively.

In the oil and gas industry, orphaned wells are defined as a class of unplugged wells whose owner/operator is unknown. Thus, other than agencies from the government, no one is responsible for the well-plugging operations and site restoration processes[1]. While some orphaned wells are well-documented with detailed information, such as name, location, and drilling details, many others lack important information and are referred to as undocumented orphaned wells. Based on a recent report from the U.S. Geological Survey (USGS), there are only 117,672 documented orphaned oil and gas wells in the 27 states in the U.S[2]. On the other hand, the Interstate Oil and Gas Compact Commission (IOGCC) reported that there are between 310,000 to 800,000 undocumented orphan wells in the 32 states of the U.S. that produce the most oil and gas, as of 2020[3]. However, it is believed that the actual number of undocumented orphan wells is much larger. Orphaned wells often present numerous environmental and health risks, including emitting methane, releasing hazardous air pollutants, creating a risk of explosion, leaking continent into underground water[4–6]. For example, according to a technical report from U.S. Environmental Protection Agency (EPA) and a recent study[4,5], in the U.S., the methane emissions from all abandoned oil and gas wells amounted to about 3% of those from natural gas and petroleum systems. However, the "documented" orphaned wells that are covered by the Bipartisan Infrastructure Law (BIL) only emit approximately 3% to 6% of total U.S. methane from all abandoned oil and gas wells. Therefore, it is necessary to find vital information on the orphaned wells such as well locations and depths for subsequent treatments to mitigate these environmental risks.

Oil and gas regulatory agencies in the U.S. maintain regulatory records (e.g., permitting documents) for wells under their jurisdiction that often contain valuable information about the location and the construction of wells. These historical records are often decades old and exist in a variety of formats that sometimes include digital PDFs but are usually scanned images or paper copies. The current practice for extracting information from historical documents related to orphaned wells involves hiring individuals to review and enter the data into a computer. This manual process requires some domain knowledge to accurately interpret the documents and correct errors, which are frequently compounded by the presence of stamps and various information formats (e.g., 45°25'28.56" and 56.358599 degrees, when dealing with unit of latitude). Given the high number of orphaned wells, it is neither practical nor realistic to manually extract and digitize this information from historical well documents. That is because the manual extraction process is labor-intensive and time-consuming.

[1]Earth & Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87544, USA. [2]Geological and Environmental Systems Directorate, National Energy Technology Laboratory, Pittsburgh, PA 15236, USA. ✉email: m.ma@lanl.gov

Therefore, it is crucial to develop an automatic information extraction workflow to analyze those historical well documents, facilitating the rapid and precise identification of the wells' location and depth information. To deal with this challenge, we developed an information extraction workflow combining text conversion techniques, (e.g., Optical Character Recognition or OCR) and large language models (LLMs). Specifically, OCR technology is used to convert different types of well documents such as PDFs and scanned images, into machine-encoded texts, which are editable and searchable data[7,8]. Next, we employed publicly available pre-trained LLMs to perform the well information extraction, during which, the converted texts are used as inputs for a properly-designed prompt. This developed workflow is based on the strong capabilities of LLMs.

LLMs are often referred to as pre-trained language models based on vast amounts of data[9,10]. Recently, artificial intelligence and machine learning have rapidly advanced and been widely adopted in the geoscience and subsurface flow fields for various applications. These include well control/production optimization in oil/gas applications[11,12], reconstruction of complex spatial fields for geospatial analysis[13], for upscaling geomechanical properties[14], for geological CO_2 storage modeling[15,16], for rapid forecasting and history matching in unconventional reservoirs[17], and for inference of random medium properties[18]. As one type of artificial intelligence model, LLMs can be described as extensive, pre-trained statistical language models that utilize neural networks[19]. The development and advances in LLMs are very fast. These developments include the introduction of new models and increased model parameter sizes, along with incorporating domain information for fine-tuned LLMs. New fine-tuned versions of base models are released many times per day, and new base models such as Llama, Mistral, and Mixtral are also released frequently. Currently, many LLMs are referred to transformer-based neural language models. These models typically possess billions of parameters and are trained using an extremely large dataset[19]. Due to their emergent ability and generalizability[20], LLMs are capable of generating text, understanding natural language, translating, summarizing content, and performing sentiment analysis, among other capabilities. Examples of applications of LLMs can be found in the following categories: translation[21], sentiment analysis[22], question and answering[23], code generation[24], summarization[25], and chatbots[26]. In the field of hydrology and earth science, a brief overview of opportunities, prospects, and concerns using ChatGPT was provided[27]. The research topic addressed in this work pertains to the question-answering category. In other words, we pose specific questions to the LLMs based on well records and anticipate that the LLMs will generate the desired answers, after analyzing the provided text. Our objective is to leverage LLMs' capability for processing text as an alternative approach to overcome challenges associated with the manual extraction of well information from historical documents, as highlighted above.

In this work, we mainly focused on Llama 2 family of Large Language Models. Llama 2 is an updated version of Llama 1[28] and it was trained on a mix of data that are publicly available[29]. In addition, there is a 40% increase in the pre-training corpus, with the model's the context length being doubled when compared with Llama 1. Meta's release of Llama 2 family consists of several pre-trained Llama 2 models, ranging from 7 billion to 70 billion parameters, along with their corresponding fine-tuned LLMs for dialogue use cases. Training Llama 2 models is not trivial, as they require advanced graphics processing unit (GPU) clusters. To train these models, Meta used two clusters equipped with NVIDIA A100 GPUs. It took about 3,311,616 GPU hours to train these models and with 539 tCO_2eq generated. According to Touvron et al.[29], Llama 2-chat models, in general, have a better performance than some open-source models on a series of safety and helpfulness benchmark tests. In addition to that, the authors also claimed that Llama 2 models achieve performance comparable to some closed-source models in their human evaluations. Because of their superior performance and open-source nature, Llama 2 models were used for analysis in this work.

In order to interact with LLMs and receive responses, it is common to use prompts[9]. A typical prompt consists of three elements: instruction, context, and input text[20]. As a new field of study, the goal of prompt engineering to improve LLMs performance for a given task by creating and refining prompt contents[20]. Recently, various prompting approaches have been developed to improve the reasoning capability of LLMs[30]. One of these examples is the chain-of-thought strategy proposed by Wei et al.[31], in which the LLMs are asked to provide a series of intermediate reasoning steps and to improve the final performances for complex reasoning tasks[30,31]. In this work, we optimized prompt contents including the approach of chain-of-thought in order to improve the performance of well information extraction tasks.

The contribution of this work can be summarized as follows: First, we developed a new LLM-based workflow for well information extraction. To the best of our knowledge, the use of LLMs to extract critical information relevant to managing orphaned oil and gas wells has not been widely reported in the literature. Therefore, this work could serve as an example for information identification tasks for other researchers and the research communities. Second, we conducted a detailed analysis of the impact of prompts, model sizes, and the chain-of-thought strategy on the information extraction performance. Third, the developed workflow can be easily deployed and we believe that employing this workflow can potentially accelerate information digitization from historical well documents.

The rest of this paper is organized as follows: we will first introduce our detailed methodology related to the workflow of information extraction, historical well records, text conversion, the theory of LLMs, Llama 2, and performance evaluations in Section 2. We will present the extraction results including various treatments that are incorporated in this work in Section 3, which is followed by a brief discussion of the potential impacts of this study, challenges, and the corresponding opportunities in Section 4. Finally, we will summarize the major findings and provide the potential future works in Section 5.

Materials and methods

In this section, we provide a detailed description of the proposed information extraction workflow, historical well records, large language models used for information extractions, and performance evaluation method.

Overview of the information extraction workflow via large language models (LLMs)

The proposed workflow for information extraction from orphaned well historical records is presented in Figure 1. As shown in this figure, the first step involves converting historical documents into text via text conversion approaches such as optical character recognition (OCR). Next, the converted texts are subjected to LLMs. Here, we integrate the texts into some predefined prompt templates to form the final question prompts.

After running the LLMs with the complete prompt, an answer in text format can be generated, as shown in Figure 2. The answer can be examined, and if the result is satisfactory, the information extraction task for this historical document is completed. Otherwise, we may need to refine the prompt or switch to different LLMs to achieve the desired outputs.

In this work, we used a standard for-loop to automatically extract the information of interest from the 160 well documents. It is worth noting that we used quantized Llama 2 models in this work to reduce the memory usage. Specifically, for example, for Llama 70B model, we utilized the Llama-2-70B-chat-GPTQ[32] obtained from the Hugging face[33] due to the reduced size, instead of its standard version. For the Llama-2-70B-chat-GPTQ, the GPTQ algorithm[34], was employed to quantize Llama 2 models within AutoGPTQ library.

Currently, we have access to only 160 documents from Colorado and Pennsylvania. This small dataset allowed us to quickly validate our approach. Once a larger dataset becomes available, the developed framework is expected to scale easily for information extraction. We acknowledge that this sample size, particularly for Pennsylvania, may not be sufficient for a highly reliable statistical evaluation. However, this dataset reflects the current limitations of available data for our study. We are actively working to obtain additional well documents from other states, and once these data are available, our workflow can be readily applied to extract information on a larger scale. Our scope of this study is to propose a novel process for information extraction from well records by leveraging the capabilities of LLMs and to test this concept.

The following subsections will cover the detailed methodologies for each major step in the information extraction workflow.

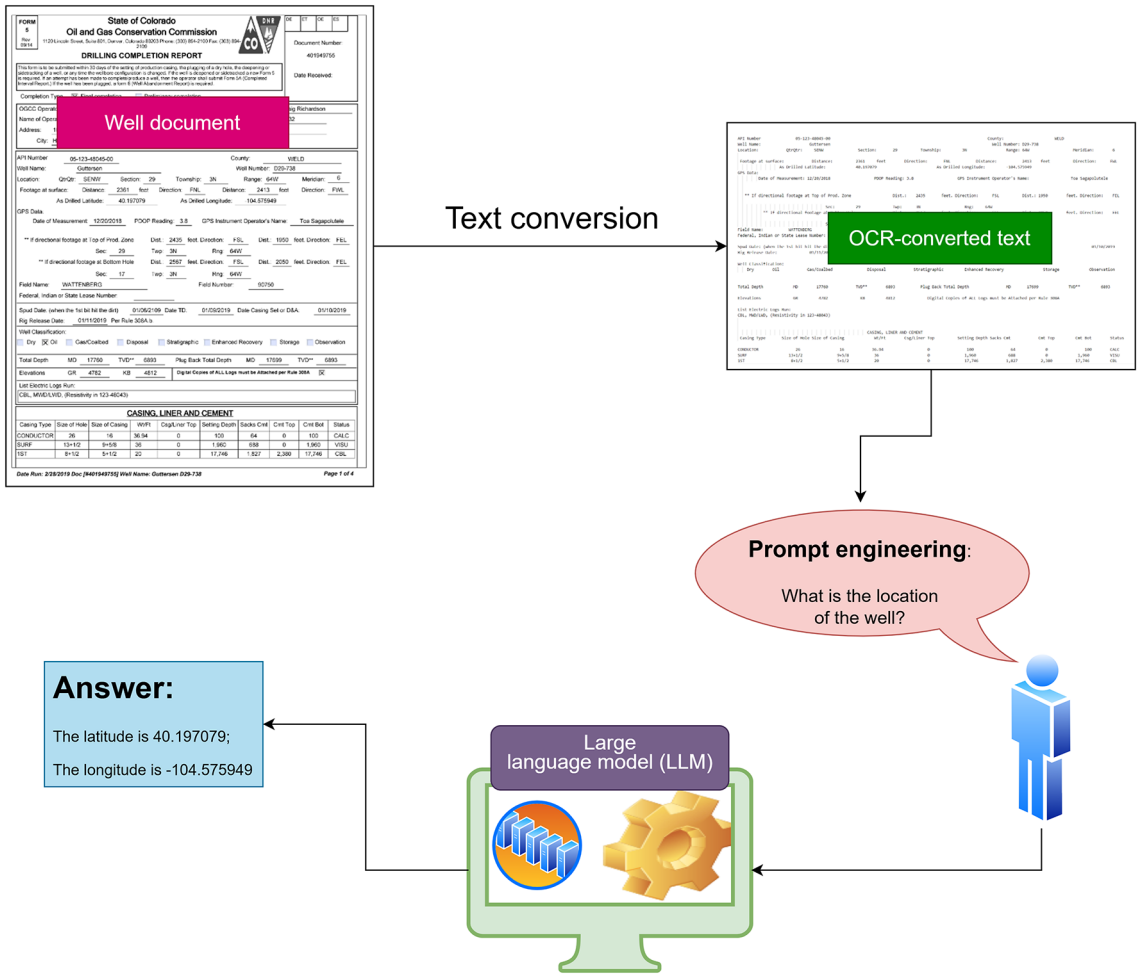


Fig. 1. The proposed workflow for well information extraction via LLM.

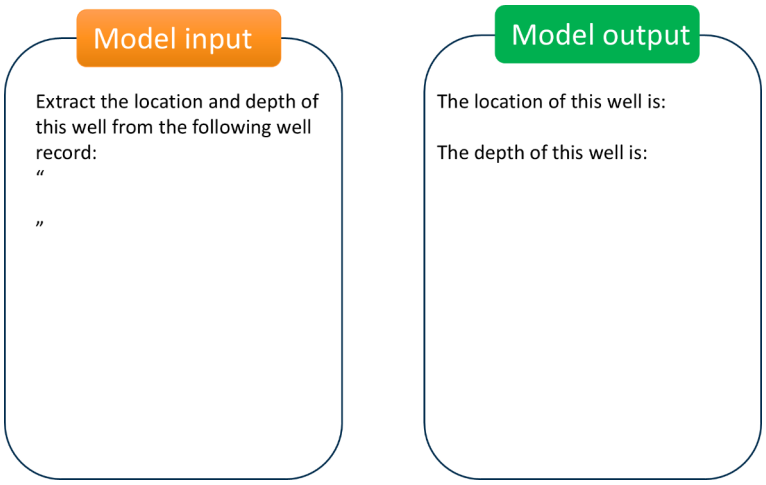
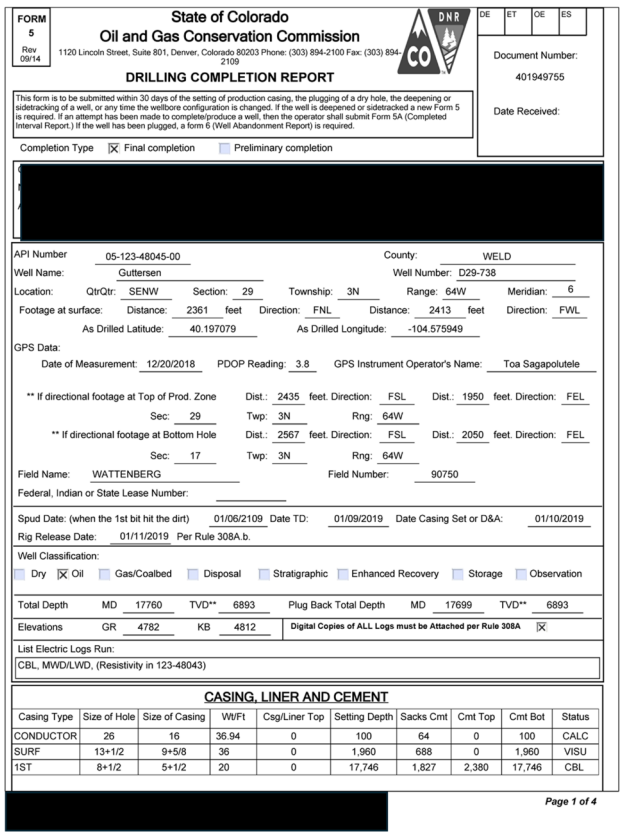
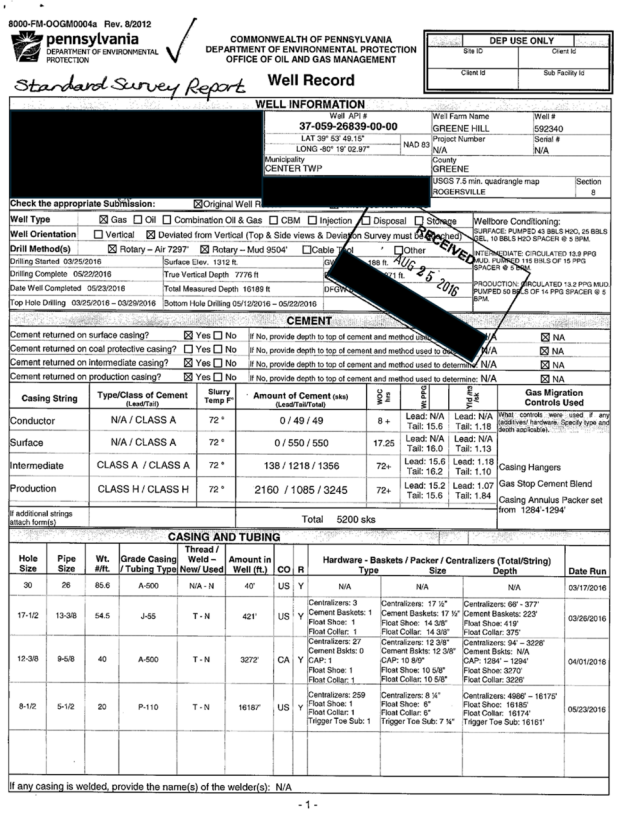


Fig. 2. An illustration of model inputs and outputs for LLM. Note that we aim to show the structure of the model's input and output. One has to provide specific well record texts to the model input section, and the LLM would generate the corresponding detailed output in terms of well location and depth.



(a) Colorado drilling completion report



(b) Pennsylvania well record

Fig. 3. Examples of well records used in this study (some sensitive information was blocked).

Historical well records

In this study, we analyzed two types of well records: well drilling completion reports from Colorado and well record reports from Pennsylvania, as illustrated in Figure 3. These types of well records are commonly utilized by oil and gas regulatory agencies to document the construction history of wells. However, it is worth noting that each jurisdiction tracks well information with their own records that have unique formats. The multitude

of oil and gas jurisdictions in the U.S. and the differences between the records they use increases the practical challenge of digitizing well information into a unified platform for characterizing orphaned wells.

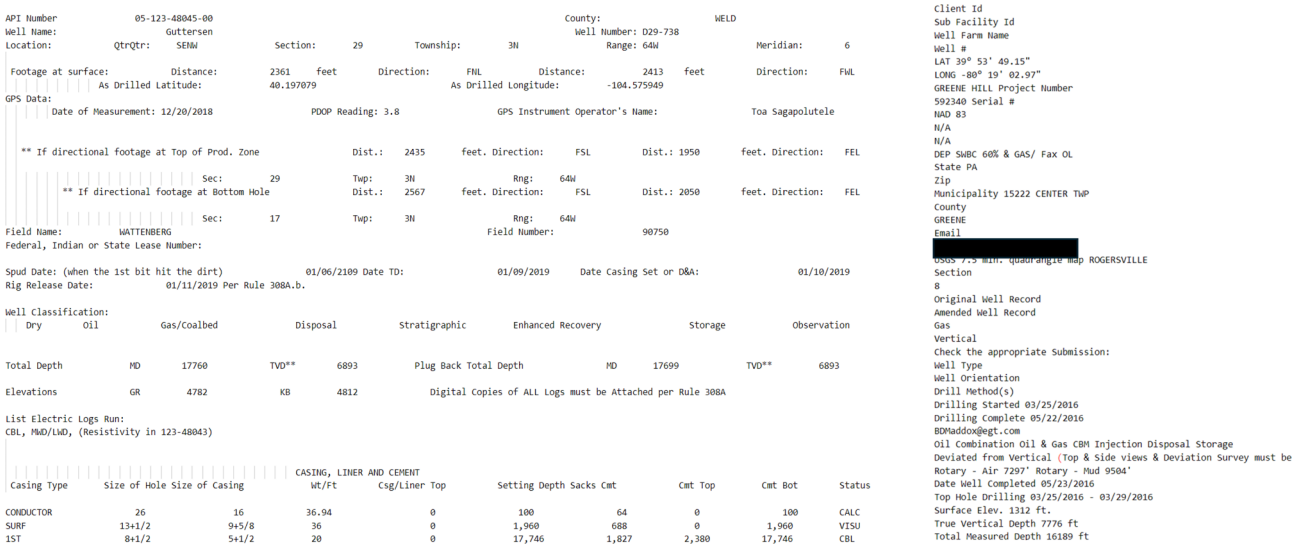
The preliminary dataset assembled for this study includes 150 well drilling completion reports from Colorado and 10 well records from Pennsylvania for demonstration purposes. The well records presented in Figure 3 contain a wealth of information, such as the operator's name, address, and phone number; the American Petroleum Institute number (a unique identifier assigned to each oil and gas well), name and location of the well; the spud date; the depth; and details on casing, liner, and cement. Although well records contain an abundance of information, the location and depth data are crucial for well remediation and will be extracted using LLMs in this work. That is because depth information provides a better understanding of the casing depth.

As shown in Figure 3, we can see that well drilling completion reports from Colorado are clean. On the other hand, the well records from Pennsylvania contain many hand-written words and stamps. For example, in the top left corner, there are three hand-written words: "Standard Survey Report". There is a stamp on the mid-right side of this record, which shows "RECEIVED AUG 25 2016". In addition, the middle part of the document is somewhat blurred with grey shadow. All these hand-written words, marks, and stamps increase the challenge of information extraction using LLMs. That is because the LLMs employed in this work require texts as input; therefore we must utilize text conversion technologies (e.g., OCR) to convert the image-based well records into text.

Text conversion

Plain text was acquired from the Colorado and Pennsylvania well records using two approaches selected based on the original format of the document: 1) PDF-to-text conversion and 2) optical character recognition (OCR). Colorado well records were stored in text-based PDF format, which enabled a direct extraction of embedded text using the open-source tool *pdftotext*[35]. Pennsylvania well records were stored as scanned image files, which have no embedded text. Consequently, Google's Enterprise OCR, made available through their Document AI API[36], was used to convert text in the Pennsylvania records into a machine-readable format.

Figure 4 displays a portion of the text information extracted from the two examples shown in Figure 3. When compared with the original documents in Figure 3, the quality of plain text conversion is acceptable, as the information presented in Figure 4 matches that in the original documents. For example, the converted information of well locations (latitude and longitude) agrees with that in the two documents in Figure 3. Another observation is that the formatting of converted information is not the same. The structure of the PDF-converted-text in Figure 4a preserves the alignment and structure of the original document as in Figure 3a. However, the OCR converted text in Figure 4b does not maintain a similar table-style structure to that in the well record shown in Figure 3b. Instead, the words within one single line in the image are divided into multiple rows in the OCR-processed text. The lack of correct structure in OCR-converted text poses a significant challenge for information extraction using LLMs as it requires LLMs to have advanced understanding capability to analyze the overall text. To improve performance, more advanced computer vision approaches should be applied for text conversion, which is beyond the scope of this paper. Once we have converted the text information, the next step is to feed it into a pre-designed prompt for LLM for extracting well location and depth.



API Number 05-123-48045-00 County: WELD
Well Name: Guttersen Well Number: D29-738
Location: QtrQtr: SENW Section: 29 Township: 3N Range: 64W Meridian: 6

Footage at surface: Distance: 2361 feet Direction: FNL Distance: 2413 feet Direction: FWL
As Drilled Latitude: 40.197079 As Drilled Longitude: -104.575949
GPS Data:
Date of Measurement: 12/20/2018 PDOP Reading: 3.8 GPS Instrument Operator's Name: Toa Sagapolutele

** If directional footage at Top of Prod. Zone Dist.: 2435 feet. Direction: FSL Dist.: 1950 feet. Direction: FEL
Sec: 29 Twp: 3N Rng: 64W
** If directional footage at Bottom Hole Dist.: 2567 feet. Direction: FSL Dist.: 2050 feet. Direction: FEL
Sec: 17 Twp: 3N Rng: 64W
Field Name: WATTENBERG Field Number: 90750
Federal, Indian or State Lease Number:

Spud Date: (when the 1st bit hit the dirt) 01/06/2109 Date TD: 01/09/2019 Date Casing Set or D&A: 01/10/2019
Rig Release Date: 01/11/2019 Per Rule 308A.b.

Well Classification:
Dry Oil Gas/Coalbed Disposal Stratigraphic Enhanced Recovery Storage Observation

Total Depth MD 17760 TVD** 6893 Plug Back Total Depth MD 17699 TVD** 6893

Elevations GR 4782 KB 4812 Digital Copies of ALL Logs must be Attached per Rule 308A

List Electric Logs Run:
CBL, MWD/LWD, (Resistivity in 123-48043)

CASING, LINER AND CEMENT

Casing Type	Size of Hole	Size of Casing	Wt/Ft	Csg/Liner Top	Setting Depth	Sacks Cmt	Cmt Top	Cmt Bot	Status
CONDUCTOR	26	16	36.94	0	100	64	0	100	CALC
SURF	13+1/2	9+5/8	36	0	1,960	688	0	1,960	VISU
1ST	8+1/2	5+1/2	20	0	17,746	1,827	2,380	17,746	CBL

Client Id
Sub Facility Id
Well Farm Name
Well #
LAT 39° 53' 49.15"
LONG -80° 19' 02.97"
GREENE HILL Project Number
592340 Serial #
NAD 83
N/A
N/A
DEP SWBC 60% & GAS/ Fax OL
State PA
Zip
Municipality 15222 CENTER TWP
County
GREENE
Email
[redacted]
USGS 7.5 Min. quadrangle map ROGERSVILLE
Section
8
Original Well Record
Amended Well Record
Gas
Vertical
Check the appropriate Submission:
Well Type
Well Orientation
Drill Method(s)
Drilling Started 03/25/2016
Drilling Complete 05/22/2016
BDMaddox@egt.com
Oil Combination Oil & Gas CBM Injection Disposal Storage
Deviated from Vertical (Top & Side views & Deviation Survey must be
Rotary - Air 7297' Rotary - Mud 9504'
Date Well Completed 05/23/2016
Top Hole Drilling 03/25/2016 - 03/29/2016
Surface Elev. 1312 ft.
True Vertical Depth 7776 ft
Total Measured Depth 16189 ft

(a) Colorado drilling completion report **(b)** Pennsylvania well record

Fig. 4. Part of the texts extracted from the two well records shown in Figure 3. Some sensitive information was blocked.

Large language models (LLMs)

LLMs are machine learning models trained on vast amounts of data. This training enables them to comprehend and produce text that closely resembles human writing. The sheer scale of these models, coupled with the large amounts of data they are trained on (on the order of trillions of tokens), allows them to learn complex patterns and relationships within the text. As the training progresses, these models develop abilities to perform a variety of tasks. For example, they can accurately answer queries, summarize vast amounts of information, and generate new text that is both coherent and contextually sound.

Llama 2[29], developed by Meta AI, is a large language model that has attracted attention from the research community for its capabilities. It follows a structure similar to GPT (Generative Pretrained Transformer)[37], which relies on stacked attention layers to process and generate text. These layers work by focusing on different parts of the input text to determine what is important and what is not. This mechanism enables Llama 2 to process and generate text effectively, understanding the context and nuances of the input text.

Available in different sizes, from smaller versions like Llama 2 7B to the largest, Llama 2 70B, these variants differ in their processing power and the depth of understanding they can provide. Larger models, while requiring more computational resources, can deliver more accurate interpretations of data. Llama 2 operates under an open-weights regime, meaning the model weights are accessible to the public, but the specific data used for training these models is not disclosed. Llama 2 comes in two main versions: Foundational and Chat. The Foundational model is a general-purpose tool for text completion, while the Chat model has been further refined with techniques like supervised fine-tuning and Reinforcement Learning from Human Feedback (RLHF) to enhance its abilities to be a useful assistant. In our work, we focused on the Chat-type models, which are optimized for tasks requiring in-depth analysis and information extraction.

As mentioned previously, prompt engineering is a crucial aspect of working with chat models. It involves crafting the input text (or *prompt*) to guide the model in generating a desired output. Through effective prompt engineering, users can steer the focus of LLMs and improve the quality of the extracted information. Moreover, LLMs like Llama 2 can perform zero-shot learning, which allows the model to make predictions or generate responses in tasks it has not explicitly been trained on. Contrarily, few-shot learning for LLMs refers to the process of a model learning from a small number of examples. For a specific task, we can provide a few demonstrative examples in the prompt to enhance the performance through few-shot learning. Another useful concept for LLM prompting is the *chain-of-thought* approach[31]. This involves the model breaking down a problem into smaller, manageable parts, similar to how humans approach complex problems. This method can enhance the model's ability to understand and solve intricate tasks, making it a very useful approach for analyzing and extracting data from extensive and complex records. In this work, we mainly tested zero-short learning and chain-of-thought methods.

LLMs performance evaluation

Although additional information is available from the well documents, in this work, we focused only on the location (latitude and longitude) and depth (true vertical depth) of each well. We employed one metric to assess the performance of our information extraction process. The metric is the accuracy based on offset or A_{OS}, which is defined as:

$$A_{OS} = \frac{N_{OS}}{N_T} \times 100\% \tag{1}$$

where A_{OS} denotes the accuracy; N_{OS} represents the number of entries that are within the offset threshold of the true value; N_T represents the total number of entries. In an ideal case, A_{OS} would be 100% if the workflow generates results that are accurate, which may not always be the case in reality. In this paper, we calculated A_{OS} for location and depth information extraction only rather than latitude, longitude, and depth. That is because the location can be represented as latitude and longitude. The location offset is calculated based on geographical distance, also known as geodetic distance, which is the shortest arc length between two locations along the Earth's surface, using GeographicLib package[38].

The rationale behind this metric is that it is also acceptable if a certain LLM generates a close approximation to the true value. For example, if the extracted location, in terms of latitude and longitude, is within 10 meters of the true location, we treat the result as correct. Similarly, if the extracted depth is within a range of 10 feet ($\sim$3 meters), we would accept this extraction result. From a practical point of view, an offset of 10 meters for the well location is considered acceptable because field operators can easily locate the well based on the extracted well location information.

Results

In this section, we demonstrate the capabilities of LLMs for information extraction using Llama 2 models. We begin by comparing the performance of Llama 2 70B model with various prompts. Next, we illustrate the performance differences among Llama 2 7B, 13B, and 70B models using the optimal prompt. Finally, we showcase the effectiveness of implementing the chain-of-thought strategy with Llama 2 70B model and compare its performance with other LLMs, again using the selected optimal prompt.

Prompt index	Prompt	Explanation
Prompt 1	Extract the location using latitude and longitude, and well depth of the well described in this well completion report. Output only the latitude, longitude, and depth in JSON format as numbers, not strings, in a clean version. Only output the JSON and nothing else. Here is the OCR'd contents of the well completion report:	This prompt directs the LLM to extract the well's location (latitude and longitude) and depth from the well completion report, and to output the numbers in JSON format. This is the simplest prompt for this task.
Prompt 2	*Extract the drilled latitude (in degrees), longitude (in degrees), and true vertical depth (TVD) of the well described in this well completion report.* Output only the latitude, longitude, and depth in JSON format as numbers, not strings, in a clean version. Only output the JSON and nothing else. Here is the OCR'd contents of the well completion report:	Provide more detailed instructions for reporting the well's location using decimal degrees, and outputting the well depth specifically in terms of True Vertical Depth (TVD).
Prompt 3	Extract the drilled latitude (in degrees), longitude (in degrees), and true vertical depth (TVD) of the well described in this well completion report. *Do not report depth in terms of Measured Depth (MD). Keep in mind that this text is extracted using optical character recognition (OCR), so the format may be jumbled. This well is in the western hemisphere, so the longitude should be negative.* Output only the latitude, longitude, and depth in JSON format as numbers, not strings, in a clean version. Only output the JSON and nothing else. Here is the OCR'd contents of the well completion report:	Additional instructions are provided to ensure the LLM not report measured depth, and the longitude should be negative given the location of the well of interest.
Prompt 4	Extract the drilled latitude (in degrees), longitude (in degrees), and true vertical depth (TVD) *(not footage at surface and not plug back total depth)* of the well described in this well completion report. Do not report depth in terms of Measured Depth (MD). Keep in mind that this text is extracted using optical character recognition (OCR), so the format may be jumbled. This well is in the western hemisphere, so the longitude should be negative. *In addition, the true vertical depth cannot be negative.* Output only the latitude, longitude, and depth in JSON format as numbers, not strings, in a clean version. Only output the JSON and nothing else. Here is the OCR'd contents of the well completion report:	Ensure the LLM not using footage at surface and plug back total depth as well as depth information. Also, ensure that well depth information from the well completion report is not negative. If a negative value is found, it should be corrected to the corresponding positive value. This is the most complicated prompt for this task.

Table 1. Different prompts for information extraction. The italicized texts in the second column highlights the differences between the current prompt and the previous one.


```
{
"latitude": 40.197079,
"longitude": -104.575949,
"depth": 6893
}
```


Fig. 5. An example of information extraction output using Llama 2 70B.

Prompt engineering for Llama 2 70B

Before applying any LLM for information extraction, we must formulate and select the optimum prompts for the Question-Answering task. This can be achieved through prompt engineering, which involves designing proper prompts to achieve desired outcomes from LLMs[39,40].

To design the best-performing prompt, we used a trial-and-error approach with iterative refinements. During this process, domain knowledge of the oil and gas industry and geosciences was applied to meet our specific data extraction requirements. For example, we specified that longitude and latitude should be in decimal format to ensure compatibility with future processing. As shown in Table 1, we designed a total of four prompts, from Prompt 1 to Prompt 4, to locate wells and identify their depths from the documents. To design these prompts, we began by manually reviewing a few representative documents to obtain a preliminary understanding of their contents and structures. Based on our domain knowledge and specific requirements, we first created a simple and straightforward Prompt 1. We then added additional constraints and guidance to subsequent prompts to enhance their focus on information extraction. For example, we directed the LLMs to extract true vertical depth as the depth information rather than measured depth (since both types are present in the documents) and enforced that longitude be negative given the wells' geographic locations. Once these prompts were designed, they were subjected to information extraction testing via LLMs, as shown in Figure 1.

Table 1 provides detailed information on four proposed prompts including prompt index, prompt content, and the corresponding explanation. Prompt 1 is the simplest one by just instructing LLMs to extract well information in terms of latitude, longitude, and depth information and to report in a JSON format. If users lack detailed information about the documents, the simplest prompt can be used directly without extensive domain knowledge. On the other hand, Prompt 4 is the most comprehensive one, ensuring that: (1) reported latitudes and longitudes are drilled latitudes and longitudes, and use decimal degrees as the unit; (2) longitudes are negative, given the well's location in the U.S.; (3) only true vertical depth is exported as depth information, despite that other depth information, e.g., measured depth, are available; (4) the true vertical depth is a positive number. By combining the proposed prompt with converted text through a text extraction process, a complete question was created for LLMs, which was then subjected to LLMs to perform information extraction. It is important to note that once the questions are formulated using the prompts, they can be directly utilized across various LLMs without any further adjustments.

After running Llama 2 70B model with a question, an output as shown in Figure 5 can be obtained. Figure 5 represents the output from Llama 2 70B model for the drilling completion report in Colorado (in Figure 3a) using Prompt 1. As expected, the output is in the format of a JSON file within the Python environment.

Prompt index	Location	Depth
Prompt 1	**100%**	48%
Prompt 2	**100%**	**100%**
Prompt 3	**100%**	**100%**
Prompt 4	**100%**	**100%**

Table 2. Information extraction results using Llama 2 70B model with different prompts for Colorado well completion reports.

Prompt index	Location	Depth
Prompt 1	60%	**90%**
Prompt 2	60%	**90%**
Prompt 3	60%	**90%**
Prompt 4	**70%**	**90%**

Table 3. Information extraction results using Llama 2 70B model with different prompts for Pennsylvania well records.

Specifically, the output for this example contains the names "latitude", "longitude", and "depth", along with their corresponding numbers as instructed by the prompt. In this example case, the latitude, longitude, and depth are "40.197079", "-104.575949", and "6893" ft, respectively. The extracted information exactly matches the true values, demonstrating the good performance of Llama 2 70B model with Prompt 1. It should be noted that although JSON-style outcomes are generated by LLMs after analyzing the text-based well records, the current workflow does not have the capability to save the outputs directly as local JSON files. Therefore, an additional post-processing step is required to save the information extraction outputs as local files.

Let's now examine the final extraction performance for the four prompts across the 160 well records in Colorado and Pennsylvania. Given the fact that the well records from the U.S. are not in the same format, we here evaluate the performance of LLMs for information extraction separately. Again as introduced in Section 2, we used A_{OS} as the metric for performance evaluation. We reiterate that we used 10 meters and $\sim$3 meters (10 ft) as the thresholds for computing well location and depth offset, respectively. For the 150 Colorado drilling well record documents, the information extraction results obtained by Llama 2 70B with four different prompts are presented in Table 2.

Clearly, excellent location extraction performances were obtained using Llama 2 70B model for Colorado cases. As shown in this table, the values of A_{OS} for location reach 100% despite the varying contents of the prompt. This observation reveals that the locations of all 150 wells were correctly extracted from the well drilling completion reports in Colorado using the LLM and proposed prompts. In terms of well depth extraction, we find that Llama 2 70B model with Prompt 1 (the simplest prompt) yielded the lowest accuracy. Specifically, the value of A_{OS} were only 48%, indicating that the locations of only 72 documents were correctly identified. In other words, Llama 2 70B model, when using Prompt 1, encountered difficulty in extracting information from the remaining 78 documents. Despite utilizing an offset of 10 ft for computing A_{OS}, the result here shows that the identified depth, using the LLM with Prompt 1, deviates by more than 10 ft from the true value for those 78 documents. For example, in one drilling well completion report, the actual well depth is 6806 ft, however, the extracted value is 17529 ft, which is about 10723 ft away from the true value. This observation illustrates that for this case, extracting well depth information was much more challenging than extracting well location information using Prompt 1. Except Prompt 1, Llama 2 70B with the remaining prompts resulted in very reliable depth extraction performance with 100% accuracy. This result demonstrates that more detailed prompts (Prompts 2 to 4) enable more reliable information extraction compared to a simple prompt (Prompt 1).

Table 3 compares well information extraction results for the 10 Pennsylvania well records using Llama 2 70B with these predefined four prompts in Table 1. The major difference between the Pennsylvania and the Colorado case studies is that Llama 2 70B provided inferior results for Pennsylvania cases. Clearly, none of the prompts resulted in completely correct extraction for the 10 Pennsylvania well records. For the location, Llama 2 70B model with the best-performing prompt (i.e., Prompt 4) resulted in an accuracy of 70%. This corresponded to 7 correctly extracted well locations. For the depth information extraction task, a value of 90% was achieved for A_{OS} from all four prompts, indicating that we obtained correct depth information for 9 out of 10 documents. Two possible reasons may explain this performance. First, the original image document contains some errors. For example, one Pennsylvania record shows a longitude of "77.670522", which should be the negative value "-77.670522". As a result, OCR also missed the "-" sign in the converted text. Despite adding, "This well is in the western hemisphere, so the longitude should be negative," in Prompts 3 and 4, Llama 2 70B may still have difficulty correcting this error. The second reason relates to unit conversion. Some Pennsylvania well records use degrees, minutes, and seconds for latitude and longitude, which differs from our request for decimal degrees. Llama 2 70B model used in this work did not perform the unit conversion completely correctly. To verify our hypothesis, we conducted another test by: (1) manually correcting any errors/issues in the OCR-converted texts for the Pennsylvania well records, given the small number of such records, and (2) re-running the information

Prompt index	Location	Depth
7B	77.33%	82.67%
13B	66%	97.33%
70B	**100%**	**100%**

Table 4. Information extraction results using three Llama 2 models with Prompt 4 for Colorado well completion report.

Prompt index	Location	Depth
7B	30%	40%
13B	**70%**	**90%**
70B	**70%**	**90%**

Table 5. Information extraction results using three Llama 2 models with Prompt 4 for Pennsylvania well records.

extraction workflow with Llama 2 7B and Prompt 4. As expected, we achieved 100% accuracy in the location extraction task, with all 10 correct location extractions from the Pennsylvania well records. This demonstrates that LLM performance can significantly improve when OCR-converted texts are more reliable. For the depth information extraction task, we observed that high accuracy was achieved with all four prompts. Specifically, the A_{OS} of all Llama 2 70B model runs reached 90%, which is very close to the accuracy obtained from the Colorado completion reports. A detailed examination of this failed record shows that it had a depth of "0" in the true vertical depth field, but Llama 2 models used the measured depth of "381" instead.

On the other hand, we observed that the extraction accuracy for locations increased with the complexity of prompts. For A_{OS}, Llama 2 70B led to 70% accuracy using Prompt 4, compared to only 60% accuracy for the remaining three prompts for the location information extraction task. The results in this case reveal that more complicated prompts result in better extraction performance. Based on the information extraction results in Tables 2–3, we see that a more detailed prompt often leads to better information extraction results. In the following sections of this paper, Prompt 4 was used for the investigation.

Comparison of different Llama 2 models

In this test, we used the best prompt via zero-shot learning from the previous section to test the extraction performance of the three Llama 2 models, i.e., 7B, 13B, and 70B. Here, Prompt 4 was employed within Llama 2 7B and 13B to extract well location and depth information from the 160 well records. Subsequently, we compared these extraction results with those from Llama 2 70B model. Tables 4-5 show the comparison results for Colorado and Pennsylvania cases, respectively. These two comparisons reveal that, in general, as the size (model parameters) of Llama 2 increases, better performance is achieved, though some deviations from the trend are observed, as shown in Table 4.

For the 150 drilling completion reports in Colorado, Llama 2 7B model achieved an accuracy of 82.67% accuracy in terms of A_{OS} for depth, which is lower than the 97.33% accuracy achieved by Llama 2 13B model. As expected, neither of the smaller models can surpass the 70B model in depth extraction. However, surprisingly, we find that for the location extraction task, the 7B model yielded a slightly better result when compared with the 13B model, contrary to expectations. For example, the Llama 13B model only achieved 66% for A_{OS}. Interestingly, the Llama 7B slightly outperformed it, achieving a higher accuracy rate 77.33% for A_{OS}. For the 10 Pennsylvania well records, we observed a consistent pattern: the larger model yielded better information extraction results. This conclusion is supported by data on both location and depth. For example, for the location information extraction, Llama 2 7B, 13B, and 70B models resulted in accuracy values of 30%, 70%, and 70%, respectively. For the depth case, Llama 2 models with 70B and 13B parameters significantly outperformed the 7B model. The results presented here demonstrate that larger LLMs are generally more effective for information extraction tasks. It is recommended that users opt for larger models if they have sufficiently powerful hardware support, as more advanced hardware is required to run larger LLMs.

Impact of the chain-of-thought on the performance of Llama 2 70B

As presented in the previous sections, Llama 2, regardless of the prompt used or the size employed, was difficult to extract completely correct information from the 10 Pennsylvania well records. In this work, we also explored the possibility of enhancing extraction performance by incorporating the chain-of-thought approach with Llama 2 70B, using Prompt 4. Given that the text extraction performance from the Colorado well drilling completion reports is reliable, we focused solely on the 10 Pennsylvania reports. We implemented the chain-of-thought approach to Prompt 4 in Table 1 by adding the following words to the end of the prompt: *Please explain your detailed steps to get the numbers*. By incorporating this strategy, Llama 2 70B model generates the following output, as presented in Figure 6.

Unlike those without the chain-of-thought strategy, the texts generated by Llama 2 here exhibited more detailed "thinking" steps for extracting the numbers for location and depth. As shown in the texts, Llama 2 first

{
"latitude": 41.02733333333334,
"longitude": -79.4538888888889,
"TVD": 6038
}

Explanation:

1. The latitude and longitude values are found in the "LAT" and "LONG" fields, respectively. The latitude value is 41°1' 56.9", which is equal to 41.02733333333334 in decimal degrees. The longitude value is 79°27' 14", which is equal to -79.4538888888889 in decimal degrees (since the longitude value is negative, it represents a location west of the prime meridian).
2. The true vertical depth (TVD) is found in the "True Vertical Depth" field, which is 6,038 ft.

Note: The "Gas Migration" and "Controls Used" fields contain information about the well's completion, but they do not provide any information about the well's location or depth. The "Casing String" and "Tubing Type" fields list the types and sizes of casing and tubing used in the well, but they do not provide any information about the well's location or depth. The "Plug-Back/Wellbore Alteration" field contains information about any plugs or fill material used in the wellbore, but it does not provide any information about the well's location or depth.</s>
PA_DEP_well_reports_image_based/37-023-20284-00-00.txt
lat = 41.02733333333334
Long = -79.4538888888889
Depth_TVD = 6038

Fig. 6. An example of LLM output for one Pennsylvania well record with implementing chain-of-thought. The true latitude, longitude, and depth are 41°1'56.9", -79°27'14", and 6038 ft, respectively.

Scenario	Location	Depth
With Chain-of-thought	60%	90%
Without Chain-of-thought	**70%**	**90%**

Table 6. Comparison of information extraction results for Pennsylvania well records using Llama 2 70B Model with and without implementing of chain-of-thought strategy.

identified the location in terms of latitude and longitude of 41°1'56.9", −79°27'14", respectively. The location numbers in degrees, minutes, and seconds accurately matched the true values in the original well record. However, as instructed by the prompt, the extracted location should be in decimal degrees, instead of degrees, minutes, and seconds. Therefore, another implicit task for the LLM is to convert the numbers from degrees, minutes, and seconds to decimal degrees, which needs a certain degree of mathematical skill. As shown in its output, Llama 2 70B directly converted the latitude of 41°1'56.9" to 41.02733333333334. However, the correct conversion should result in 41.032472. Despite the two values being very close, a slight difference remained.

The corresponding information extraction results with incorporating the chain-of-thought strategy are presented in Table 6. It is interesting to observe that, with the chain-of-thought employed, Llama 2 70B model resulted in the same or even worse extraction performance than that achieved without using the chain-of-thought strategy. Specifically, Llama 2 70B with two strategies resulted in the same level of accuracy for depth extraction. For location extraction, Llama 2 70B with the chain-of-thought strategy yielded 60% in terms of A_{OS}, which is lower than the 70% by Llama 2 70B without chain-of-thought strategy. The comparison presented here reveals that the chain-of-thought strategy offered limited improvement for location and depth information extractions for Llama 2 70B. It is anticipated that if a post-processing procedure is applied to convert the location units from degrees, minutes, and seconds to decimal degrees, the results could be potentially improved. However, this is beyond the current scope of this paper. In addition, combining LLMs with external tools through the strategy of function calling may be one potential solution to this precise mathematical problem.

Comparison with other LLMs

In this work, we also tested the information extraction workflow using three additional models, including Mixtral 8×7B, Llama 3.1 70B, and Llama 3.1 405B. For the Mixtral 8×7B model, the results are presented in Table 7. Interestingly, the Mixtral 8×7B did not yield better information extraction results compared to Llama 2 70B model used in this study.

State name	Location	Depth
Colorado	**99.33%**	**97.33%**
Pennsylvania	30%	50%

Table 7. Information extraction results using Mixtral 8×7B model with Prompt 4 for Colorado and Pennsylvania well reports.

On the other hand, the Llama 3.1 models were just released in July 2024. We applied the SambaNova Platform (https://sambanova.ai/) to implement the full Llama 3.1 70B and 405B models. Here, we used the Llama 3.1 models to extract information for the 10 Pennsylvania well records only. Our results showed an improved performance when Llama 3.1 models were used. Specifically, the values of A_{OS} for location and depth reached 80% and 90% with the Llama 3.1 70B model, and 90% and 100% with the Llama 3.1 405B model. The Llama 3.1 405B model extracted all correct depth information, including the case with a depth of "0". It made only one error in location extraction out of 10 documents in terms of A_{OS}. This reveals that the Llama 3.1 405B model has stronger mathematical capabilities for converting units of latitude and longitude. This additional result demonstrates the potential of more powerful and new models to achieve higher accuracy.

Discussions

In this section, we briefly discuss the potential impacts of the developed workflow, the current challenges, and potential opportunities for applying LLMs to tasks such as information extraction. This discussion is based on our results from extracting well location and depth data using 160 documents. In addition, given that the development of LLMs is progressing rapidly, it is possible that some of the information summarized here may not accurately reflect the latest advancements in LLMs.

Potential impacts

The developed LLM-based information extraction framework has great potential to accelerate the document digitization process. We expect that this workflow can save significant time and reduce costs for large-scale document information extraction tasks. Since the LLM-based workflow can operate continuously and in parallel, the efficiency of information extraction could be improved. Although the workflow was developed for orphaned well characterization task, it has a wide range of potential applications, as similar large-scale data extraction challenges exist across various fields.

It is worth noting that the information extraction framework may not achieve 100% accuracy in locating wells and identifying depth information from documents, as seen, for example, in the Pennsylvania case study. High accuracy is likely necessary for the practical application of this framework in some real-world scenarios, as inaccuracies could result in significant operational costs. Therefore, enhancing the accuracy of this framework as much as possible is recommended. In practice, this methodology can be combined with other techniques to more reliably identify well locations for orphaned well applications. Examples include remote sensing technologies, such as aero-magnetometers and fixed-wing drones equipped with magnetometers, as reported by O'Malley et al.[41].

Enhance text conversion quality from historical documents

As introduced previously, the current information extraction tasks require that the original historical documents be converted to texts before feeding into LLMs. This is because the LLMs employed in this work are designed to process textual inputs. Thus, the tasks heavily depend on the accuracy of the text conversion process used (e.g., OCR). However, even the best text conversion techniques still struggle to achieve 100% accurate text conversions from documents such as PDFs and images. To deal with this challenge, it is recommended to further advance text extraction techniques to improve the accuracy and quality. Integrating computer vision techniques or machine learning algorithms could be a potential area. An alternative path to improving text extraction quality is to utilize large multi-modal models that can extract textual information from the images directly. This is a promising direction for future research.

Improve the capabilities of LLMs

The technology of LLMs advances rapidly in terms of new models, increased parameter sizes, and capabilities[42]. However, given that numerous LLMs are available from both the private and public domains, exploring other LLMs is necessary to get a better result. In this paper, we mainly focused on testing Llama 2 models. As discussed in Section 3, Llama 2, despite the use of various prompts, changes in model parameter sizes, and the incorporation of the chain-of-thought strategy, could not achieve precisely correct information extraction from historical well documents. Therefore, it is worth testing other LLMs for the same task.

Many commercial LLM-based tools are available for document processing and information extraction, including Amazon Textract, OpenAI's GPT-4, and Google Document AI. For instance, we used Google Document AI's Enterprise OCR for the text conversion task in this study. Generally, these commercial models or tools deliver better performance than some open-source models, likely due to their larger model sizes. However, commercial LLMs have certain limitations compared to smaller open-source models. First, due to their commercial nature, users must pay for access, which can increase costs. Second, proprietary models like GPT-4 come with potential data security concerns. For example, some cloud-based tools require documents to be uploaded to the cloud for processing. These requirements may limit their applicability in certain industrial

and governmental contexts. Smaller open-source models, such as Llama 2 70B, can offer a balanced solution, providing cost-effective and more secure options for various tasks. Specifically, once downloaded, they can be used directly on appropriate hardware without incurring additional costs related to utilization of the models. Additionally, they can operate offline and locally, without the need to upload data to a cloud environment. Given the rapid development and advancement of LLMs, it is expected that more advanced open-source LLMs will become available to the public.

Another opportunity lies in fine-tuning the pre-trained LLMs for specific tasks. In this work, we focused solely on the zero-shot learning strategy, without performing any fine-tuning. However, fine-tuning LLMs could potentially be a better option, if feasible. Currently, we are investigating the improvement of fine-tuned LLMs for information extraction and will report their findings on performance in a future publication. Furthermore, it is also possible to incorporate large multi-modal models for information extraction. Specifically, these models can directly take images or PDFs as inputs, eliminating the need for text conversion using OCR techniques. Although not employed in this study, it is also advisable to implement some post-processing procedures to enhance the information extraction performance.

Overcome the challenges of extreme hardware requirements

In order to use these LLMs offline, we must meet the hardware requirements, especially regarding GPUs due to the extremely large size of the LLMs. For example, according to the Hugging Face data repository, the total size of the standard version of Llama 2 70B-chat-hf is approximately 280 GB. Additionally, Hugging Face suggests using 4 $\times$ NVIDIA A100 GPUs for the deployment of Llama 2 70B models. While using the pre-trained LLMs as presented in this paper does not demand extensive computational resources, it still requires higher-end GPUs to run. In the information extraction task, we utilized an NVIDIA RTX A6000 GPU with 48 GB of memory. Even with this GPU, we encountered difficulties loading the full standard version of Llama 2 70B model. This was the reason for using the quantized Llama 2 models in this study. Specifically, when we applied 4-bit quantization to the Llama-2-70B-chat-GPTQ model, the GPU memory usage was approximately 42 GB according to Stoelinga[43], which fit within the available memory of an NVIDIA RTX A6000. Therefore, the extreme hardware requirements may hinder the wide applications of LLMs. One way to address this challenge is through using more advanced GPUs. Given the recent rapid advancements in GPU technology, the situation should continue to improve. Additionally, the development of smaller LLMs could also be a viable solution. Another alternative is to use commercial LLMs that are only available through an API if costs and data security are not concerns.

Concluding remarks

In this work, we presented an LLM-based workflow to extract vital information from well records for orphaned well management, including the well's location and depth. Extracting data from historic records is currently a time-consuming and costly process. The information contained in well records is critically important for successful plugging operations to reduce environmental impacts such as methane leakage from wellbores. To demonstrate the capability of information analysis workflow, we primarily focused on Llama 2 models, which are publicly available. To facilitate information extraction, we developed multiple prompts, varying the instructions from the simplest to the most complex one. Different variants of Llama 2 were also evaluated, including the 7B, 13B, and 70B models. Additionally, we also employed the chain-of-thought approach in an attempt to enhance performance. We tested the developed workflow using a dataset of 160 well records. Although this number is quite small, the goal of this paper is to prove the concept of this method. We emphasize that these forms are used only for validation of the approach, not for training the models. The development of an information extraction framework capable of handling much larger datasets of well documented information is an ongoing project.

Several major conclusions can be drawn from the results. First, the content of the prompt impacts the final extraction results, even when an identical LLM is used. In this work, we found that Llama 2 70B model with Prompt 4 led to the best performance. The general trend is that the information extraction performance improves with the complexity of the prompt instructions. Therefore, it is recommended to optimize prompt content before using LLMs. Second, the size of the model is an important parameter that influences the result. With Llama 2 models, better performance was often obtained when a larger model was used. For example, Llama 2 70B model outperformed the smaller models, including the 7B and 13B variants. Third, although Llama 2 models achieved 100% accuracy for the Colorado reports, they still had difficulties in correctly extracting information from some Pennsylvania well reports. For example, Llama 2 70B extracted the correct location information in the units of degrees, minutes, and seconds after incorporating a chain-of-thought strategy, but it did not accurately convert it into decimal degrees as instructed.

While the developed workflow achieved good performance, especially for the PDF-based documents, opportunities for further improvement still remain. These include: (1) improving the quality of text conversion from historical documents, since the current workflow relies heavily on that; (2) fine-tuning the pre-trained LLMs for this specific task using a smaller dataset; (3) executing these information extraction tasks on higher-end hardware to enhance the results; (4) utilizing large multi-modal models that can directly process PDFs and images, thereby eliminating the need for text extraction; and (5) utilization of LLM function calling techniques to aid the LLM with routine tasks like converting. These techniques could automate significant portions of the extraction workflow, accelerating the plugging of abandoned wells and enabling large-scale data collection for research purposes.

Open research section

The well documents analyzed in this manuscript are publicly available. Specifically, Colorado records were acquired from the Colorado Energy and Carbon Management Commission's online system (COGIS): ECMC

Data (state.co.us), and Pennsylvania records were acquired from the Pennsylvania Geological Survey's EDWIN online tool: Home - EDWIN Subscriptions (pa.gov). We have permission to use these well records for the analysis in this paper.

Received: 21 May 2024; Accepted: 29 November 2024
Published online: 30 December 2024

References

1. Boutot, J., Peltz, A. S., McVay, R. & Kang, M. Documented orphaned oil and gas wells across the United States. *Environmental Science & Technology.* **56**, 14228–14236 (2022).
2. Merrill, M. D., Grove, C. A., Gianoutsos, N. J. & Freeman, P. A. Analysis of the United States documented unplugged orphaned oil and gas well dataset. *Technical Report from US Geological Survey.* (2023).
3. IOGCC. Idle and orphan oil and gas wells: State and provincial regulatory strategies 2021. *Technical Report from Interstate Oil and Gas Compact Commission (IOGCC)* (2021).
4. EPA. Inventory of U.S. greenhouse gas emissions and sinks: 1990-2020. *Technical Report from United States Environmental Protection Agency (EPA)* (2022).
5. Kang, M. et al. Environmental risks and opportunities of orphaned oil and gas wells in the United States. *Environmental Research Letters.* **18**, 074012. https://doi.org/10.1088/1748-9326/acdae7 (2023).
6. Raimi, D., Krupnick, A. J., Shah, J.-S. & Thompson, A. Decommissioning orphaned and abandoned oil and gas wells: New estimates and cost drivers. *Environ. Sci. Technol.* **55**, 10224–10230. https://doi.org/10.1021/acs.est.1c02234 (2021).
7. Eikvil, L. Optical character recognition. *citeseer. ist. psu. edu/142042. html* **26** (1993).
8. Chaudhuri, A. et al. *Optical character recognition systems.* (Springer, 2017).
9. Chang, Y. et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology.* (2023).
10. Topsakal, O. & Akinci, T. C. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. *International Conference on Applied Engineering and Natural Sciences.* **1**, 1050–1056 (2023).
11. Ma, Z., Kim, Y. D., Volkov, O. & Durlofsky, L. J. Optimization of subsurface flow operations using a dynamic proxy strategy. *Mathematical Geosciences.* **54**, 1261–1287. https://doi.org/10.1007/s11004-022-10020-2 (2022).
12. Ma, Z. & Leung, J. Y. Design of warm solvent injection processes for heterogeneous heavy oil reservoirs: A hybrid workflow of multi-objective optimization and proxy models. *Journal of Petroleum Science and Engineering.* **191**, 107186. https://doi.org/10.1016/j.petrol.2020.107186 (2020).
13. Santos, J. E. et al. Development of the senseiver for efficient field reconstruction from sparse observations. *Nature Machine Intelligence.* **5**, 1317–1325 (2023).
14. Zhang, B., Ma, Z., Zheng, D., Chalaturnyk, R. J. & Boisvert, J. Upscaling shear strength of heterogeneous oil sands with interbedded shales using artificial neural network. *SPE Journal.* **28**, 737–753 (2023).
15. Yan, B., Harp, D. R., Chen, B. & Pawar, R. J. Improving deep learning performance for predicting large-scale geological co 2 sequestration modeling through feature coarsening. *Scientific Reports.* **12**, 20667 (2022).
16. Ma, Z., Guo, Q., Viswanathan, H., Pawar, R. & Chen, B., Deep Learning Assisted History Matching and Forecasting: Applied to the Illinois Basin—Decatur Project (IBDP). Available at SSRN: https://ssrn.com/abstract=5019810 or http://dx.doi.org/10.2139/ssrn.5019810 (2024).
17. Srinivasan, S. et al. A machine learning framework for rapid forecasting and history matching in unconventional reservoirs. *Scientific Reports.* **11**, 21730. https://doi.org/10.1038/s41598-021-01023-w (2021).
18. Gao, K. & Modrak, R. T. Machine learning inference of random medium properties. *IEEE Transactions on Geoscience and Remote Sensing.* **62**, 1–13. https://doi.org/10.1109/TGRS.2024.3367541 (2024).
19. Minaee, S. et al. Large language models: A survey. arXiv preprint arXiv:2402.06196 (2024).
20. Pan, S. et al. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* (2024).
21. Koshkin, R., Sudoh, K. & Nakamura, S. Transllama: Llm-based simultaneous translation system. arXiv:2402.04636 (2024).
22. Sun, X. et al. Sentiment analysis through llm negotiations. arXiv:2311.01876 (2023).
23. Zhuang, Y., Yu, Y., Wang, K., Sun, H. & Zhang, C. Toolqa: A dataset for llm question answering with external tools. In Oh, A. et al. (eds.) *Advances in Neural Information Processing Systems.*, vol. 36, 50117–50143 (Curran Associates, Inc., 2023).
24. Wang, Y., Wang, W., Joty, S. & Hoi, S. C. H. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. arXiv:2109.00859 (2021). 2109.00859.
25. Shekhar, S. et al. Towards optimizing the costs of llm usage. arXiv:2402.01742 (2024).
26. Tan, T. F. et al. Fine-tuning large language model (llm) artificial intelligence chatbots in ophthalmology and llm-based evaluation using GPT-4. arXiv:2402.10083 (2024).
27. Foroumandi, E. et al. ChatGPT in hydrology and earth sciences: Opportunities, prospects, and concerns. *Water Resources Research.* **59**, e2023WR036288. https://doi.org/10.1029/2023WR036288 (2023).
28. Touvron, H. et al. Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023).
29. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
30. Huang, J. et al. Large language models can self-improve. arXiv preprint arXiv:2210.11610 (2022).
31. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems.* **35**, 24824–24837 (2022).
32. TheBloke. Llama 2 70b chat - gptq (2024).
33. Wolf, T. et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019).
34. Frantar, E., Ashkboos, S., Hoefler, T. & Alistarh, D. GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323 (2022).
35. Pdftotext. pdftotext - portable document format (pdf) to text converter (version 3.00). *software available at* https://linux.die.net/man/1/pdftotext (2024).
36. Google. Google Document AI. *online tool available* https://cloud.google.com/document-ai?hl=en (2024).
37. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. Improving language understanding by generative pre-training. *OpenAI* (2018).
38. Karney, C. Geographiclib. *online at* (2015).
39. Liu, V. & Chilton, L. B. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.*, 1–23 (2022).
40. Reynolds, L. & McDonell, K. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7 (2021).
41. O'Malley, D. et al. Unlocking solutions: Innovative approaches to identifying and mitigating the environmental impacts of undocumented orphan wells in the united states. *Environmental Science & Technology.* **58**, 44. https://doi.org/10.1021/acs.est.4c02069 (2024).

42. Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. Science in the age of large language models. *Nature Reviews Physics*. 1–4 (2023).
43. Substratus, S. Calculating gpu memory for large language models (2023).

Acknowledgements
This work is supported by the U.S. Department of Energy's Undocumented Orphan Well program through the CATALOG consortium (catalog.energy.gov). LA-UR number "LA-UR-24-23837".

Author contributions
Zhiwei Ma contributed to Methodology, Formal Analysis, Investigation, Writing and Reviewing the Original Draft, and Visualization; Javier E. Santos Contributed to Editing and Reviewing the Original Draft; Greg Lackey Contributed to Data Collection and Editing and Reviewing the Original Draft; Hari Viswanathan Contributed to Reviewing the Original Draft, Project Supervision, and Funding Acquisition; Daniel O'Malley Contributed to Data Collection, Methodology, Result Discussion, Reviewing the Original Draft, Project Supervision, and Funding Acquisition.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Z.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024